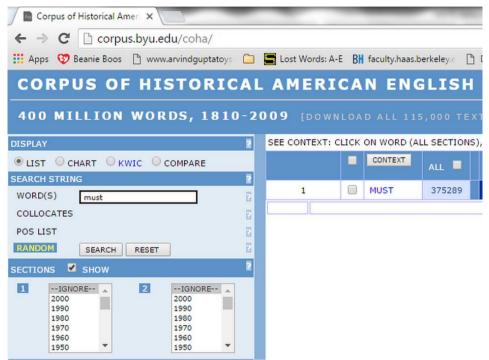


Supplementary Document

3.1

The purpose of this section is to describe and explain the procedures utilized in my study on the frequency of modals and semi-modals in the THC and COHA corpora. Enough detail has been supplied to explain how the procedures were carried out.



Note the URL to access the COHA corpus. The level of access I needed required the creation of a free acount. Anyone can apply for, and receive permission to, open an account.

Next, I wanted to collecy frequency data for both modals and semi-modals on a historical basis. For this aspect of the search to work, the bottom-left box (Sorting and Limits) has to have Frequency highlighted under Sorting, and the Frequency needs to be set to 10 years. The next graphic shows how to set up these parameters for COHA. Again, this process can be replicated by anyone wishing to verify my analyses.

	CORPUS OF HISTORICA
	400 MILLION WORDS, 1810-2
	DISPLAY 2
	● LIST ○ CHART ○ KWIC ○ COMPARE
	SEARCH STRING
	WORD(S) must
	COLLOCATES
	POS LIST
	RANDOM SEARCH RESET
	SECTIONS 🗹 SHOW
	1 IGNORE 2 IGNORE 2000 1990 1990 1990 1990 1980 1980 1980 1980 1970 1970 1960 1960 1950 ▼ 1950 ▼
	SORTING AND LIMITS
	SORTING FREQUENCY V
	MINIMUM FREQUENCY V 10
	CLICK TO SEE OPTIONS
Now let's exam	nine the result:

CORPUS OF HISTORICAL AMERICAN ENGLISH																					
400	400 MILLION WORDS, 1810-2009 [DOWNLOAD ALL 115,000 TEXTS]																				
DISPLAY	SEE CONTEXT: CL	ICK ON WORD (ALL SEC	TIONS), NUMBE	R (ONE SE	CTION),	OR [CONTE	EXT] (SELE	CT) [HEL	P]											CON	MPARE V
● LIST		CONTEXT	ALL 🔲	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
CHART	1	MUST	375289	2198	8589	15693	19159	19727	20631	22730	23945	22195	24927	26438	24783	22788	21954	20638	19251	17603	15675

The graphic above can't be seen well, because it is horizontally ruled, so let's zoom in:

🖉 🔤 Corp	ous of Historical Ameri	×		-		-	-
$\leftarrow \ \Rightarrow$	C 🗋 corpus.by	u.edu	ı/coha/				
Apps	🐯 Beanie Boos 📋	www.ar	vindguptatoys (🗀 🔚 Lost Wo	rds: A-E	BH faculty	y.haas.berke
COR	PUS OF H	IS	TORIC	AL AME	RIC	AN E	INGL
400	MILLION W	ORD	S, 1810	- 2009 [D	OWNLO	AD AL	L 115,0
DISPLA	SEE CONTEXT: CLIC	K ON W	ORD (ALL SECT	IONS), NUMBER	(ONE SEC	CTION), C	OR [CONTE
● LIST		-	CONTEXT	ALL 💻	1810	1820	1830
CHART	1		MUST	375289	2198	8589	15693
о кwic							
COMPAR							
SEARCH STRING							

The key point here is to note that, for each word in the search, the automated results return the frequency of that word for every 10 years in COHA. For example, in the search for *must*, the graphic above shows that there were 2198 uses of this word in 1810, 8589 in 1820, and 15693 in 1830. In fact, given the parameters I entered earlier, COHA returned a frequency for the word *must* in every decade from 1810 to 2000. Note that these results are for raw frequency. In order to ensure more valid comparisons, I delimited my results to words per million, which accounted for the different size of the virtual corpus in different years.



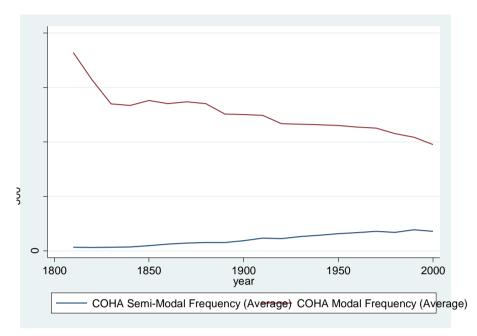
What these figures allowed me to do is to perform some statistical analyses on how the frequencies of *must* (and the other modals and semi-modals in my analysis) changed over time.

So now these data can be entered in a program for statistical analysis. I used Stata for this purpose, but I used also eViews for one of the analyses.

File Edit Data Graphics Statistics User Wir	ndow Help							
	0							
(R)	StataCorp 4905 Lakeway Drive	-						
Special Edition	College Station, Te: 800-STATA-PC 979-696-4600 979-696-4601 (fax)	xas 77845 USA http://www.stata.com stata@stata.com						
Single-user Stata perpetual license: Serial number: 401406214766 Licensed to: User User1								
Notes: 1. Unicode is supported; see h 2. Maximum number of variables		help set_maxvar.						

Note that, for the variable names in Stata and eViews, THC refers to the Hansard Corpus, while COHA is the Corpus of Historical American English. The names of the variables can be more easily understood once the prefixes are udnerstood, but there are some special terms that require explanation. Note that I created average (semimodal_avg and modal_avg) variables by averaging out all of the modals and semi-modals in the THC and COHA corpora. This procedure allowed me to trace the frequency changes of all modals and semi-modals, which was an important consideration when trying to understand the higher-level trends in frequency changes. Of course, I analyzed the frequency change of individual modals and semi-modals are variable for both modals and semi-modals let me achieve insights that couldn't have been generated by analyses of the individual modals and semi-modals.

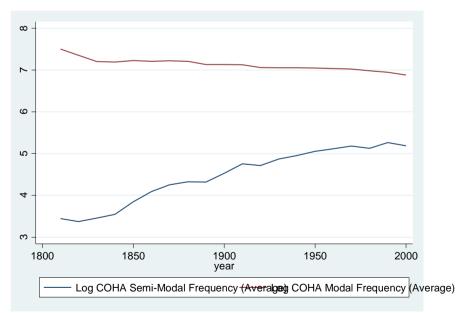
Note also that the ln prefix refers to a logarithmic transformation. Because raw frequencies were quite different (in terms of magnitude) between modals and semi-modals, I needed a method to compare them on the same graph. Using a log-transform procedure allowed me to achieve this kind of comparative insight, which was an important part of the analysis. Consider the graph below:



Also, note that the syntax to reproduce this graphic in Stata is as follows:

tsline coha_semimodal_avg coha_modal_avg

Note that the raw count of modal frequencies is much higher, so it isn't really possible to see the convergence of modals and semi-modals. Now, when log-transformed data are used, the result is as follows:



Note that the syntax to reproduce this graphic in Stata is as follows:

tsline lncoha_semimodal_avg lncoha_modal_avg

What takes place in the log-transformed model is a much easier illustration of the convergence of the frequency of modals and semi-modals in COHA over time. This analysis tells us that, in COHA, the ratio of semi-modals to modals has been increased over time. Following Smith's substitution thesis, I wanted to see if the relationship between modal decline and semi-modal increase in COHA could be quantified through linear regression. In Stata, the command for this analysis coha modal avg

20 45.91 0.0000 0.7183 0.7027

104.81

[95% Conf. Interval]

would be reg coha_modal_av coha_semimodal_avg, and the results are as follows:

Coef.

Source	SS	df	MS	Number of obs	
				F(1, 18)	
Model	504309.655	1	504309.655	Prob > F	
Residual	197736.128	18	10985.3404	R-squared	
	Second State Practice of the Constant of the			Adj R-squared	
Total	702045.783	19	36949.778	Root MSE	

. reg coha modal avg coha semimodal avg

t

P>|t|

Std. Err.

T I created several histograms in Stata for inclusion in the thesis. The dialog box below demonstrates how histograms are created in Stata.

Main	if∕in	Weights	Density plots	Add plots	Y axis	X axis	Titles	Legend	Overall	By	
Dat Var	a able:		7	ata are cont ata are discr							
Bins		Width	er of bins of bins limit of first bin			© Fr © Fr	ensity action equency ercent				
	Bar prope		vhen by() is spe	cified		_		abels to ba			

I used eViews for modeling the effect (as an impulse repsonse function, or IRF) of large changes in semi-modals on the subsequent frequencies of modals. I perfomed this analysis for hoth TCH and COHA. Note that, for the IRF to work in eViews, the 2 variables of interest (in this case, modal and semi-modals averages for COHA for every decade from 1810 to 2000) have to be opened as part of a vector autogression, or VAR:

VAR Type	Endogenous Variables			
 Unrestricted VAR Vector Error Correction Bayesian VAR 	coha_semimodal_avg coha_modal_avg			
Estimation Sample	Lag Intervals for Endogenous:			
1810 2000	12			
	Exogenous Variables			
	c			

Once the variables are opened as a VAR, then the IRF can be generated as follows:

Impulse Responses	×
Display Impulse Definition	
Display Format	Display Information Impulses:
 Multiple Graphs Combined Graphs 	coha_semimodal_avg
Response Standard Errors	Responses:
Analytic (asymptotic)	
Monte Carlo Repetitions: 100	Periods: 10
	Accumulated Responses
	OK Cancel

In this dialog box, semimodal average is the independent (impulse) variable, and modsl average is the response (dependent) variable. Once OK is clicked, the model generates the graphs I used in my paper.

The same procedures that I used for COHA were used for THC as well. First, let me show how the Hansard Corpus can be retrieved:

corpus.byu.edu corpora, size, queries = better resources, more insight

Created by Mark Davies, BYU. Overview, search types, looking at variation, corpus-based resources, upcoming.

The most widely used online corpora -- more than 130,000 distinct researchers, teachers, and students each mo

English	# words	language/dialect	time period
Wikipedia Corpus (with virtual corpora)	1.9 billion	English	-2014
Global Web-Based English (GloWbE)	1.9 billion	20 countries	2012-13
Corpus of Contemporary American English (COCA)	520 million	American	1990 -2015
Corpus of Historical American English (COHA)	400 million	American	1810-2009
TIME Magazine Corpus	100 million	American	1923-2006
Corpus of American Soap Operas	100 million	American	2001-2012
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s
Hansard Corpus (British Parliament)	1.6 billion	British	1803-2005

BYU has a corpus site where both COHA and TCH can be accessed. Note that, in order to pull up modal or semi-modal frequencies by decade, the same procedure is followed for THC as was followed for COHA (see screen shot on following page). Again, the key with this form of analysis is to choose 10 for the Frequency, and to highlight the Frequency box. Once these steps are taken, then the resulting readout sorts the search word into frequencies per decade. Now that, as with COHA, I configured the THC results to be frequency per million words, which allowed better comparisons from year to year (which is an important consideration for all virtual corpora, given that the size of each corpus changes from year to year). Raw frequencies are therefore not an appropriate input for analysis.

HANSARD CORPUS (BRI
7.6 MILLION SPEECHES,	1.6 B
DISPLAY	2
● LIST ○ CHART ○ KWIC ○ COMPARE	
SEARCH STRING	2
WORD(S) must	2
COLLOCATES	2
POS LIST	2
SEMANTIC CATEGORIES WORDS	2
RANDOM SEARCH RESET	2
SHOW DECADE SPEAKER	2
1 IGNORE 2 IGNORE 2000 1990 1990 1990 1990 1980 1980 1980 1980 1970 1960 1950 ▼	•
SORTING AND LIMITS	
SORTING FREQUENCY	2
MINIMUM FREQUENCY V 10	2
CLICK TO SEE OPTIONS	2

Finally, here are partial screenshots of my Stata dataset:

thc_must[25]						
	year	thc_would	thc_will	thc_can	thc_could	thc_may
1	1810	7035.75	803.01	664.95	3683.55	612.1
2	1820	7626.87	857.87	665.94	3591.37	601.19
3	1830	8576.56	956.85	505.13	3181.62	502.90
4	1840	7740.32	1408.42	801.02	2835.05	761.19
5	1850	7536.76	1583.71	861.57	2533.69	900.64
6	1860	7481.59	1904.69	980.07	2314.56	1026.33
7	1870	7978.34	1693.96	854.22	2529.6	910.98
8	1880	7115.66	2654.35	1179.25	2068.7	1073.44
9	1890	5638.17	3954.82	1778.64	1486.36	1435.60
10	1900	5926.54	3688.26	1776.79	1749.73	1261.88
11	1910	3762.8	5544.18	2869.16	903.37	2056.
12	1920	3577.49	5322.01	3055.82	926.35	2001.93
13	1930	3825.44	5589.6	3095.88	940	2045.75
14	1940	3854.65	6043.58	3324.96	1055.09	2047.20
15	1950	3881.74	5598.87	3243.51	1134.3	1919.5
16	1960	4230.43	5928	3146.74	1230.96	1851.3
17	1970	3971.55	6303.06	2991.04	1217.75	1758.
18	1980	3714.47	6617.56	2886.8	1126.66	1567.1
19	1990	3574.61	6985.66	2832.59	1079.62	1558.84
20	2000	3709.58	6876.96	2881.58	1141.02	1509.0

Note that all of the raw data for the study have been presented in full in the body of the study itself.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).